Text Analysis Using Automated Language Translators

CDT John Stanford
MAJ Ian McCulloh
Department of Mathematical Sciences
United States Military Academy
West Point, New York 10996

Abstract

Text analysis is a new tool with many interesting possibilities for intelligence-gathering. Software being developed by Carnegie Mellon University can output a mental model of a text with the top six concepts in that text. This can be used to automatically analyze thousands of texts to search for keywords, find trends over time, or compare two different geographic areas. The problem is that most of the texts that intelligence analysts would use are not in English. Machine translators like the Forward Area Language Converter (FALCon) produce English text that is hard for the average person to read but this machine-translated text appears to be just as useful for text analysis as human-translated text. Using machines instead of humans to translate text can save intelligence agencies time and money.

1. Introduction

The information age has brought a rapid expansion of the mass media all over the world. The media in strategically significant regions of the world contain a wealth of useful information about the attitudes and ideologies of the people in the region. Text analysis is a process that provides an automated way to process large volumes of text in order to obtain quantitative measures that describe the ideological terrain of the population or person of interest.

However, there is a problem of language translation in this kind of analysis. It would be very expensive and time-consuming to hire human translators to translate texts written in Arabic, Urdu, or any other local language into English for analysis. The alternative is machine translators like the Forward Area Language Converter (FALCon), being developed by the Army Research Labs. These programs input foreign language text and output English text. The English output can be very crude and hard to understand for most people and this leads many people to the conclusion that this software is useless (McCulloh, 2006). However, the English text that these machine translators output may be just as useful for text analysis as human-translated texts.

Text analysis can provide useful, fast, and quantitative information about the content of a given text. This paper's hypothesis is that machine-translated text is just as useful for text analysis as human-translated text.

2. Literature Review

2.1 Theoretical Framework. The theoretical aspect of this type of analysis rests upon the assumption that text can be represented as a network of concepts (Sowa, 1984). This network is represented quantitatively in a matrix called an adjacency matrix. If we define a list of concepts in the text, $\{c_1, c_2, c_3, ...c_n\}$, the adjacency matrix takes the following form.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2006		**00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Text Analysis Using Automated Language Translators	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Military Academy,Department of Mathematical Sciences,West Point,NY,10996**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Proceedings of the 14th Annual Army Research Lab - US Military Academy Technical Symposium, Aberdeen, MD 1 Nov 2006

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**11**	

$$A = \begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3} & \cdots & r_{1,n} \\ r_{2,1} & r_{2,2} & r_{2,3} & \cdots & r_{2,n} \\ r_{3,1} & r_{3,2} & r_{3,3} & \cdots & r_{3,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{n,1} & r_{n,2} & r_{n,3} & \cdots & r_{n,n} \end{bmatrix} \qquad (1)$$

In this expression $r_{i,j}$ is a positive integer that represents the strength of the relationship between c_i and c_j.

2.2 Conversion from Text to Adjacency Matrix. This provides a very simple framework for representing the concept network of the text, but leaves us with the issue of converting the text file into this form. This is done by a beta program called AutoMap, created by researchers at Carnegie Mellon University (Carley and Diesner, 2004). To give an overview of the basics of how this program operates, it begins by preprocessing the texts. This includes a number of operations that the program performs to eliminate some of the insignificant words in the text in order to make it easier to analyze. One operation is called stemming. This cuts off parts of the word that are part of the syntax of the language, but not part of the concept (Carley and Diesner, 2004). For example, it will make plural words singular by cutting off the "s" at the end of the word. Another preprocessing step is deletion. In this step, AutoMap uses a list of words that serve a syntactical purpose in the language, but don't carry any real meaning (Carley and Diesner, 2004). This would cut out words such as "the", "is", "and", etc. The third important step is generalization. This step uses a thesaurus file to find words that are similar in meaning and replaces them all with a word that represents this group of related words so that they show up as the same concept in the text network (Carley and Diesner, 2004). For example, "America", "United States", and "USA" would all be replaced with "united_states" in the generalization step.

Once preprocessing is done, the text is ready to be analyzed. There are two parameters for this analysis that are important for this study. The first is directionality. The analysis can be done either uni-directionally or bi-directionally. In a uni-directional analysis, the relationship between two words depends on the order of the words. In a bi-directional analysis, the relationship between two words depends only on the distance between them, not the direction (Carley and Diesner, 2004). The second parameter of interest is the window size. This setting determines the distance between words that are considered connected (Carley and Diesner, 2004). Once these parameters are set, AutoMap can analyze the input texts and output adjacency matrices in XML format.

2.3 Analysis on the Concept Network. Once AutoMap outputs the matrices for the texts, another beta program developed by Carnegie Mellon University called ORA (Organizational Risk Analyzer) is used to analyze the networks. The analysis begins with loading meta-matrices in the form of XML files output by AutoMap (Carley and Reminga, 2004). There is a 'Generate Reports' command in ORA that brings up a dialog box with several options for reports that can be generated. The report that calculates the network measure relevant to text analysis is the mental model report. This report

computes the communicative power of each concept in the network and returns a list of the concepts with the highest communicative power. The communicative power is basically a measure of how central and connected a concept is in the text. Example output from ORA is shown below.

```
Concepts with High Communicative Power

Symbols (high degree centrality, high connectivity, high strength)
Number of concepts in this class = 5
1     energi                    0.5415
2     united_states             0.1369
3     technolog                 0.1120
4     plan                      0.1075
5     increas                   0.0766
```

Figure 1. Mental Model Output from ORA.

This network comes from a radio address given by President Bush on 23 Feb 2002 about his energy policy (Bush, 2002). In the address, he outlines his plan to conserve energy and invest in energy technology. This is reflected in the concepts with high communicative power. The analysis identified energy as the most significant concept in the text. Technology also appears close to the top, reflecting the president's emphasis on investing in technology. Notice that the word "energy" appears as "energi" in the network. This is an example of the effects of the stemming preprocessing step in AutoMap. In the original text, "technology" and "technological" would both have been generalized to "technolog" by the stemming feature.

3. Additional Software

There were two additional pieces of software used in this study that were developed as an ad hoc solution to problems that arose. The programs are called TextCleanup, used as an additional pre-processing step, and NTA_Script, used to automate the ORA analysis of the AutoMap output.

3.1 TextCleanup. This program does some additional pre-processing on the text before it goes to AutoMap. This program was created to allow extra flexibility in the pre-processing step that ORA doesn't have. For example, when FALCon translates an Arabic document and there are two possible translations for a word, FALCon will output both word separated by a forward-slash, e.g. 'making/report'. In the text analysis, this would show up as one concept. Rather than allowing this to happen, it's better to pick one of the words. Since there is no way of knowing which word represents the Arabic word more accurately, the first word in the pair is arbitrarily chosen. TextCleanup does this by reading in words and cutting off the part of the word after the forward-slash.

Another example of a problem solved by this program is words in the text with extra characters at the beginning or end of the word. For example, if the word 'Iraq' shows up in the text alone, then appears again at the end of a quotation, AutoMap will recognize iraq and iraq" as separate concepts. TextCleanup fixes this problem by deleting special characters such as quotation marks, semicolons, and commas.

Since the texts were all downloaded from the internet, the address of the site shows up in the file even though it's not part of the text proper. TextCleanup fixes this by removing any words that contain the character sequence 'http'.

For further details on this program, see the comments in the source code in Appendix B. As stated earlier, this code was written as an ad hoc solution to several problems that arose and is not as refined as it could be. A similar program for use by operational intelligence units should be developed by a better-qualified, professional software engineer.

3.2 NTA_Script. AutoMap outputs a meta-matrix for each text and these meta-matrices must all be processed by ORA. Unfortunately, when a communicative power analysis is run on multiple meta-matrices in ORA, output is not generated for each individual meta-matrix, but for the union of all of them. In order to get output for each individual meta-matrix, it is necessary to run them one at a time in ORA. This is very tedious, especially as the size of the data set increases.

One way around this problem is the scripting feature of ORA (Reminga, 2006). ORA is capable of running scripts executed from a command line. The script is an XML file containing the location and name of the meta-matrix file, the type of report to be run, and the location to output the report. The command line contains the filename of the script, the location of an XML file that ORA needs, and the location to output the log file. NTA_Script requires a list of the meta-matrices to be analyzed, the meta-matrices themselves, and a file containing information on the formatting of the meta-matrix files.

NTA_Script starts by opening a batch file and writing the commands necessary to set the current directory to the directory were ORA.exe is located. Next, NTA_Script opens the list of meta-matrix filenames and starts reading them in one at a time. For each file, the program writes an ORA script for that file and a new line to the batch file that runs that script. A sample line of the batch file is shown below.

```
pra.exe -script d:\ORA_Fi~1\0000000009.xml -measures d:\ORA_Fi~1\ora_xml_measures.xml -log d:\ORA_Fi~1\oraLogFile.txt
```

Figure 2. Sample Line of the Batch File.

With the scripts and the batch file to run them all written, NTA_Script has what it needs to run the analysis. The program runs the batch file and allows ORA to output the communicative power reports in *.csv format. Once it's finished with the analysis, NTA_Script cleans up after itself by deleting all the script files. It also deletes the local communicative power reports since only the global ones will be used.

Once this is all complete, the reports for each text are in their own files. The reports must all be compiled into one file to be useful. The format of the compiled file will depend on the aspect of the reports being analyzed. Like the ORA analysis, this would be extremely tedious to do manually, so NTA_Script does it automatically. It outputs two files. The first file contains a list of the six concepts from each human-translated text with the highest communicative power along with the corresponding communicative power from the machine-translated text for the same concepts. The second file simply contains the top six concepts from each human-translated text along with the top six concepts from the machine-translated text for each text.

4. Radio Address Study

4.1 Reason for Using BLUFOR Data. All the data used in this study is BLUFOR data. It all comes from US government websites. The reason for this is because friendly data is much easier to collect and the analysis process is the same as it would be for terrorist data (Graham, 2006). BLUFOR data is used to test the method. Once the method is proven, it can be applied just as easily to terrorist data.

4.2 Radio Address Study. This short study motivates the usefulness of text analysis. Every week the President delivers a radio address to the nation. It is hypothesized that text analysis data from these radio addresses over a period of time will show a trend that is consistent with actual world events. The data set consists of 94 radio addresses delivered by President Bush between September 11th and the first few weeks of Operation Iraqi Freedom (Bush, 2001-03). This first radio address in the data set is from 15 September 2001 and the last one is from 21 June 2003.

The radio addresses were run through several preprocessing steps and AutoMap was used to generate the networks. In AutoMap, the directionality was set to bi-directional and a window size of five was used. ORA generated the mental models for each of these networks. The data from each network was output to a single file resulting in 94 files full of communicative power statistics. A short program was used to compile all these files into one file with the top three concepts from each text. Part of this file is shown below.

Source File	Concept 1	Communicative Power	Concept 2	Communicative Power	Concept 3	Communicative Power
20010915.txt	violence	0.2693	united_states	0.1923	nation	0.1444
20010922.txt	united_states	0.3163	work	0.1529	economi	0.1125
20010929.txt	violence	0.3156	united_states	0.1355	nation	0.1135
20011006.txt	violence	0.2594	peopl	0.2473	united_states	0.2155
20011013.txt	united_states	0.2945	violence	0.2582	afghanistan	0.1014
20011020.txt	violence	0.2128	peopl	0.1831	world	0.1576
20011027.txt	violence	0.2766	new	0.2122	law	0.1521
20011103.txt	anthrax	0.3222	violence	0.1474	mail	0.1

Table 1. Communicative Power Data from Radio Addresses.

With this data, a timeline is made starting with 15 September 2001 and ending with 21 June 2003. A large, gray dot on the timeline indicates a radio address where the 'violence' concept is the concept with the highest communicative power. A smaller dot indicates that 'violence' has the second highest communicative power in that address. An even smaller dot indicates that it has the third highest communicative power. The timeline is shown below.

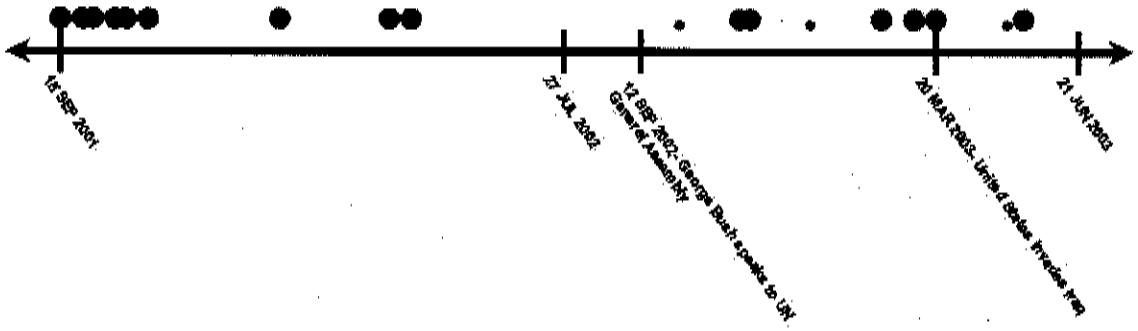

Figure 3. Timeline Plot of 'violence' Concept Occurrence.

Two key dates are also marked on the timeline. The first is 12 September 2002. This is the day when President Bush addressed the UN General Assembly and first laid out his case against Iraq. The second date is 20 March 2003, the date when the United States invaded Iraq and Operation Iraqi Freedom began. There is a cluster of 'violence' concept occurrences towards the beginning of the timeline in the aftermath of the September 11th attacks and the subsequent anthrax attacks. There are sporadic occurrences before President Bush started making the case for the war in Iraq. After 12 September 2002, there is a noticeable increase in occurrences as the President begins preparing for OIF.

4.3 Conclusions. The small study shows that text analysis data can be useful for identifying trends in BLUFOR data. The only thing that prevents this exact same method from being applied to REDFOR data is the language barrier. Now that the applications of this type of analysis have been demonstrated, the language barrier problem is addressed.

5. Machine-Translated Arabic Texts

5.1 Overview. This study also uses BLUFOR data for the same reason that BLUFOR data was used for the study on Presidential radio addresses. The data comes from the State Departments news site (US Dept of State, 2006). The State Department publishes news articles on this site and translates them into relevant languages. Most of the articles dealing with the Middle East are translated into Arabic. The original English versions of 22 articles were downloaded along with the version that was translated by a human from English to Arabic. A program called CyberTrans, which is part of the FALCon package, was used to translate the Arabic documents back into English (Swam, 1999). With the two versions for each article, a text analysis was run on all of them and compared the results by article.

5.2 Example of Translated Text. As mentioned earlier, the text translated by the machine is very messy and incoherent. Most humans would have to work hard to get even a rough idea of what the text is about. An example of machine-translated text is compared to human translated text below.

```
Army Major General Richard Zahner told reporters at the Combined Press
Information Center in Baghdad, Iraq, September 27 that, as part of the
coalition's strategy for success in Iraq, "we're having to block Shiite
extremists from linking with Iran."
```

Figure 4. Original English Text Sample.

```
informed the float/general in the military Richard ZANR the correspondents
from the station the information your joint in the capital the Iraqi Baghdad
day 27 september current groan? part from strategic/strategy the alliance
for the achievement help in Iraq? "loss forced prevention the extremists/extreme
Shiites from the attachment with Iran."
```

Figure 5. Machine-Translated English Sample.

5.3 Analysis of Data. The text analysis on this set of texts was run the same way that the analysis of the radio address data was run. The full results of this analysis are shown in Appendix A. In 16 of the 22 articles the top concept was the same for both the human and machine-translated texts. In all of the articles, the top concept in the human-translated text is one of the top three concepts for the machine-translated text. Of all the top six concepts for the human-translated texts, 69.7% are also in the top six machine-translated concepts.

Another thing to notice is that when the machine and human-translated texts differ in the concepts that they identify, the concepts that come from the machine-translated text usually describe the subject of the article just as well as the concepts that come from the human-translated text. For example, in the analysis of the 26 September 2006 article about Condoleezza Rice seeing a struggle between extremism and moderation in the Middle East, the machine-translated text analysis identifies 'iran' as one of the top six concepts while the human-translated text analysis does not. In the article, Rice talks about several extremist nations and groups in the Middle East and focuses on Iran. This is something that should probably be recognized by text analysis. This means that even though the machine-translated analysis may be different from the human-translated analysis, it does not mean that the human-translated analysis is necessarily better.

6. Conclusions and Recommendations

From the results of the text analysis on the Arabic documents, it appears that machine-translated texts are just as useful as human-translated texts for text analysis. This has significant implications for intelligence-gathering agencies. If one of these agencies wants to look for trends in terrorist publications over a period of time, they can use a computer program to automatically translate and run text analysis on a very large volume of texts instead of hiring human translators that will take much more time and cost much more money. When you hire human translators, there is also the potential for security issues that don't arise with machine-translation.

An intelligence agency could have an archive of thousands of terrorist publications over several years, run a text analysis that takes no more than a few minutes on all the documents, and have a database with the top three concepts in each text based on communicative power. If the agency wanted to graphically analyze the terrorist group's interest in weapons of mass destruction over a period of time, they could create a chart like figure 1 showing occurrences of this concept over time.

Another interesting possibility would be to create one of these charts for texts talking about IEDs and include actual IED attacks on the chart to see how strong of a relationships exists between the publication talking about IEDs and the rate of actual real-world IED attacks. If a strong correlation exists, then text analysis data from the publication could be used to predict the likelihood and frequency of attacks and drive

tactical decisions. For example, if a strong relationship is found, text analysis data from newspapers published in one town might suggest that IED attacks are more likely there than in another town. Therefore, tactical planners could decide to route convoys through the safer town.

It is recommended that intelligence analysts who have access to large volumes of REDFOR data conduct similar test analyses to confirm that this process works with that data as well as with BLUFOR data. In addition, while FALCon output may not be easily intelligible to humans, FALCon can be developed further and expanded to other languages such as Farsi so that it can be used for text analysis.

7. References

Bush, George. (2002). "President Focuses on Energy Security in Radio Address." Washington DC: Office of the Press Secretary. Available from <http://www.whitehouse.gov/news/releases/2002/02/20020223.html>.

Bush, George. (2001-03). "President Bush's Radio Addresses by date and topic." Washington, DC: Office of the Press Secretary. Available from < http://www.whitehouse.gov/news/radio/index.html>.

Carley, Kathleen and Diesner, Jana. (2004). Revealing Social Structure from Texts: Meta-Matrix Text Analysis as a novel method for Network Text Analysis. *Causal Mapping for Information Systems and Technology Research: Approaches, Advances, and Illustrations.*, Harrisburg, PA: Idea Group Publishing.

Carley, Kathleen and Diesner, Jana. (January 2004). AutoMap1.2 - Extract, analyze, represent, and compare mental models from texts.

Carley, Kathleen and Reminga, Jeffrey. (2004). ORA: Organization Risk Analyzer.

Graham, John. Class Discussion. PL497: Seminar in Behavioral Science. US Military Academy, West Point, NY. October 2006.

McCulloh Ian A., Morton, J., Jantzi, J., Rodriguez, A., Graham, J. (2006) *Efficacy in Automated Language Translators.* Proceedings of the 25th Army Science Conference. Orlando, FL. November 2006.

Reminga, Jeffrey. "RE: Communicative Power report." Email to John Stanford. 19 Oct 2006.

Sowa, J.F. (1984). *Conceptual Structures: Information Processing in Mind and Machine.* Reading, MA: Addison-Wesley.

Swam, Kathlen. (1999). *FALCon: Evaluation of OCR and Machine Translation Paradigms.* U.S. Army Research Laboratory, Aberdeen Proving Ground, MD.

US Dept of State. (2006). "News from Washington." Washington, DC: Office of the Press Secretary. Available from < http://usinfo.state.gov/usinfo/products/washfile.html>.

Appendix A: Data from Arabic Text Analysis

The following appendix identifies the top six concepts as determined from the human translation and the FALCon machine translation for each of the 22 articles in the Arabic Text Study.

Text	Top Human Concepts	Top Machine Concepts
20060926 President To Declassify Intelligence Report on Iraq	violence	violence
	media	media
	iraq	united_states
	president_bush	iraq
	united_states	president_bush
	state	talk
20060926 Rice Sees Struggle of Extremism Moderation in Middle East	power	power
	state	state
	moder	moder
	think	rice
	rice	iran
	tell	minist
20060927 Bush Releases Intelligence Report Findings on Terrorism Iraq	violence	violence
	iraq	media
	state	iraq
	media	united_states
	peopl	state
	alqaida	expans
20060927 Rice Says Time Is Running Out for Iran To Suspend Enrichment	iran	iran
	state	state
	program	program
	council	punishment
	rice	rice
	punishment	secur
20060927 Terrorists Targeting Iraqis During Holy Month of Ramadan	iraq	iraq
	caldwel	state
	power	violence
	violence	baghdad
	baghdad	alliane
	state	group
20060928 Defense Department Outlines Links Between Iraq Iran	iran	iran
	iraq	iraq
	zahner	militari
	power	power
	state	haughti
	jam	present
20060928 Hezbollah an Octopus with Tentacles Around World Officials Say	hezbollah	hezbollah
	violence	violence
	frank_urbancic	united_states
	support	support
	state	frank_urbancic
	group	talk
20060929 Tribal Leaders in Western Iraq Are Working Against Al-Qaida	iraq	iraq

	power	sean_macfarlan d
	sean_macfarlan d	united_states
	ramadi	ramadi
	tribal	polic
	alqaida	alqaida
20060929 U.S. Agency Announces $7.8 Million To Revitalize Iraqi Agriculture	iraq	agricultur
	agricultur	iraq
	expans	united_states
	usda	expans
	train	ministri
	develop	develop
20060930 Bush Says Iraq Action Has Not Worsened Terror Threat	violence	violence
	iraq	iraq
	nation	united_states
	united_states	president_bush
	new	nation
	presid	evalu
20061001 Defeating Insurgents in Iraq Will Take Lengthy Effort Rumsfeld Says	iraq	united_states
	donald_rumsfeld	iraq
	violence	donald_rumsfeld
	state	state
	united_states	violence
	take	time
20061001 Newly Passed Act Will Help Halt Iran's Weapons Program Bush Says	iran	iran
	punishment	united_states
	president_bush	punishment
	bill	bill
	program	president_bush
	state	program
20061002 Rice Seeks To Rally Moderate Forces in Middle East	state	state
	rice	moder
	middle_east	rice
	power	power
	moder	particip
	region	democraci
20061003 Rice Urges Egypt To Lead Middle East Democratization	state	democraci
	egypt	state
	democraci	united_states
	united_states	egypt
	rice	rice
	secretari	nuclear
20061003 Rice Urges Hamas To Join International Consensus for Peace	state	palestinian
	palestinian	state
	hama	united_states
	rice	intern
	intern	hama
	territori	foreign
20061004 Muslim Americans Prepare for Eid-ul-Fitr	islam	united_states
	eid_ul_fitr	celebr

	united_states	islam
	celebr	origin
	new	front
	observ	majida
20061005 NATO Commanding International Security Operations in Afghanistan	afghanistan	afghanistan
	nato	power
	power	alliance
	state	united_states
	october	state
	command	oper
20061005 Rice Makes Unannounced Visit to Baghdad	state	state
	iraq	iraq
	rice	rice
	process	minist
	secretari	oper
	reconcili	round
20061005 U.S. Says Sudan Is Intimidating Troop-Contributing Countries	sudan	united_nations
	united_nations	sudan
	darfur	darfur
	council	secur
	state	united_states
	secur	protect
20061006 Coalition Helps Iraqis Minimize Sectarian Influences on Police	polic	polic
	iraq	group
	power	train
	peterson	peterson
	train	iraq
	ministri	ministri
20061006 Rice Discusses Iraqi National Unity with Kurdish Leader	iraq	iraq
	state	kurdish
	kurdish	rice
	rice	oper
	region	state
	barzani	barzani
20061006 Sudan Gives Up Threats to U.N. Peacekeeping Contributors	sudan	state
	state	sudan
	united_nations	united_states
	power	intern
	darfur	darfur
	united_states	united_nations